Text+: Sprach- und textbasierte Forschungsdateninfrastruktur (Text+)

Sprecher: Prof. Dr. Erhard Hinrichs, Leibniz-Institut für Deutsche Sprache, hinrichs@ids-mannheim.de

Adressierte Forschungsgebiete: Text+ beabsichtigt den Aufbau einer auf Sprach- und Textdaten ausgerichteten Forschungsdateninfrastruktur, die sich zunächst auf digitale Sammlungen, lexikalische Ressourcen und Editionen konzentriert. Diese Datendomänen haben eine lange Tradition in der geisteswissenschaftlichen Forschung. Sie sind mit ausgereiften methodologischen Paradigmen verknüpft, die jeweils charakteristische, aber auch bereichsübergreifende Praktiken der Datenerzeugung, -nutzung, -analyse, -vernetzung und -kuratierung erfordern. Sie sind unabdingbar für eine breite Palette von Fachdisziplinen einschließlich, aber nicht beschränkt auf Linguistik, Literaturwissenschaft, Philologien auch der sog. 'Kleinen Fächer', Philosophie sowie sprach- und textbasierte Forschung in den Sozialwissenschaften und der Politikwissenschaft. Die drei Datendomänen sind außerdem grundlegend für interdisziplinäre Forschungspraktiken der Hermeneutik, Paläographie, Genealogie, Editionsphilologie, Lexikographie und Computerphilologie sowie Computerlinguistik.

Art der Daten: Der Name Text+ soll vermitteln, dass sich diese Initiative auf typischerweise textbasierte digitale Forschungsdaten konzentriert, die bzgl. Sprachräumen (auch über Europa hinaus) und Modalitäten von Sprache und Schriftsystemen heterogen sind; das +-Zeichen weist darauf hin, dass sprachbasierte Ressourcen auch Ressourcen und Werkzeuge für gesprochene Sprache und für multimodale Daten umfassen. Die zunächst von Text+ adressierten Forschungsdaten gliedern sich in drei Domänen:

Sprach- und textbasierte Sammlungen umfassen Sammlungen geschriebener, gesprochener oder gebärdeter Sprache und Texte sowie sprach- und textbezogene Experimental- oder Messdaten, die auf Grundlage wissenschaftlicher Kriterien gesammelt wurden. Dazu gehören: Textsammlungen (z. B. von literarischen Texten, Sachtexten, Zeitungs- und Zeitschriftentexten, Interviews, Inschriften, Handschriften, Drucken), mono- und multimodale Aufnahmen z. B. von spontaner und formaler Sprache (z. B. von Reden, Dialogen, Nachrichten, Interviews, Interaktion im Alltag), Sensordaten (z. B. EEG, Eyetracking, Artikulographie), Befragungen, Reaktionszeiten etc.

Lexikalische Ressourcen sind Daten, die die Verwendung von Wörtern in Sätzen, Texten und multimodaler Kommunikation beschreiben, darunter: Wörterbücher (mehrsprachige Wörterbücher, historische Wörterbücher, Fachwörterbücher), Enzyklopädien, Normdaten, terminologische Datenbanken, Ontologien, Wortlisten, Wortkarten und linguistische Atlanten, Übersetzungswörterbücher (für menschliche oder maschinelle Übersetzung) etc.

Editionen sind kritische Repräsentationen historischer Dokumente, wie sie in der geisteswissenschaftlichen Forschung und darüber hinaus verwendet werden. Sie bestehen aus der zuverlässigen methodengeleiteten Bewahrung, Präsentation und Kommentierung aller Arten von Texten in verschiedenen Sprachen und Schriftsystemen. Unter der Vielzahl editorischer Modelle finden sich dokumentarische oder diplomatische Editionen, Editionen zur Entstehungsgeschichte von Dokumenten und historisch-kritische Editionen.

Aus der Fülle der adressierten Forschungsdaten ergibt sich die Notwendigkeit, auf Basis transparenter, breit angelegter wissenschaftsgeleiteter Auswahlkriterien das Portfolio von Text+ initial zu definieren und kontinuierlich auszuweiten.

Datenmanagement-Maßnahmen und Services für das Forschungsgebiet: Alle Datenmanagement-Maßnahmen haben zum Ziel, Forschende bei der Erzeugung, Nutzung, Analyse, Vernetzung und Kuratierung von Forschungsdaten zu unterstützen und deren Bewahrung und Wiederverwendbarkeit, auch für die maschinelle Verarbeitung, sicherzustellen. Zertifizierte Datenrepositorien der antragstellenden und beteiligten Institutionen und perspektivisch weiterer Partner spielen eine entscheidende Rolle bei der Implementierung der Datenhosting- und Archivierungsdienste von Text+; jede Institution übernimmt dabei die Verantwortung für diejenigen Forschungsdaten, die in besonderem Maße mit ihrem Auftrag und ihrem Fachwissen in Zusammenhang stehen. Forschende werden von Beginn an bei der Erstellung von Datenmanagementplänen, der Wahl passender Repositorien und Lizenzen, in der Auswahl und Nutzung fachspezifischer Werkzeuge und Verfahren (z. B. Annotationsverfahren, Text-Data-Mining) sowie der Anwendung und Weiterentwicklung von Standards und Normdaten für die Repräsentation von Forschungsdaten und ihrer Metadaten zur Herstellung von Interoperabilität unterstützt. Auch die FAIRification von Bestandsdaten gehört zum Portfolio. In jährlichen Community-basierten Review- und Innovations-Zyklen wird das Daten-Portfolio ebenso wie das Angebot an digitalen Methoden und Werkzeugen weiterentwickelt.

Spezielle Anforderungen an das Forschungsdatenmanagement: Der Umgang mit der Vielfalt und Breite sprach- und textbasierter Daten stellt besondere Anforderungen an das Forschungsdatenmanagement. Diese Vielfalt schließt unterschiedliche Metadatenformate ebenso ein wie unterschiedliche Datenformate und unterschiedliche Grade von Strukturiertheit der Daten. Aufgrund von Urheberrechts- und Datenschutzanforderungen verteilen sich diese heterogenen Forschungsdaten auf viele verschiedene Akteure und müssen zu einem großen Teil an unterschiedlichen geographischen Standorten angeboten werden. Daraus ergibt sich die Notwendigkeit einer ortsverteilten Forschungsdateninfrastruktur, die den FAIR-Prinzipien entspricht und gleichzeitig den rechtlichen Anforderungen genügt und sich dabei an gängigen Standards und Empfehlungen, wie z.B. dem BSI-Grundschutz und der ISO 27001, orientiert. Text+ widmet sich der Orchestrierung bestehender und zukünftiger Aktivitäten, die sich über den gesamten Forschungsdatenlebenszyklus erstrecken.

Erfahrungen und Hintergrund im Datenmanagement: Die antragstellenden und beteiligten Institutionen verfügen über beträchtliche Erfahrungen beim Aufbau und Betrieb eines ortsverteilten Netzwerkes von zertifizierten Datenzentren. Die Zertifizierung, die eine verbindliche Anforderung für alle an Text+ beteiligten Datenrepositorien darstellen wird, erfolgt nach internationalen Standards des Core Trust Seals. Bei der Umsetzung der FAIR-Prinzipien kann Text+ auf die folgenden bestehenden Komponenten eines Forschungsdatenmanagements zurückgreifen: Einhaltung von Metadatenstandards, Anwendung von standardkonformen Protokollen für das Metadaten-Harvesting, Zuweisung von persistenten Identifikatoren für Forschungsdaten, Anwendung standardkonformer Protokolle zur Authentifizierung, Autorisierung und Identifikation (AAI) sowie für das Harvesting. Die antragstellenden und beteiligten Institutionen von Text+ haben ausgereifte Datenspeicherlösungen entwickelt, die bereits seit einigen Jahren in Betrieb sind und die für die Kuration und Archivierung sprach- und textbasierter Forschungsdaten genutzt werden.

Relevante (internationale) Partner und vorhandene Infrastrukturen: Die antragstellenden und beteiligten Institutionen von Text+ repräsentieren die Interessensgruppen, die Forschungsdaten für die Geisteswissenschaften bereitstellen und nutzen: Hochschulen, Wissenschaftliche Bibliotheken, Datenzentren der Digital Humanities, Mitglieder der Union der

deutschen Akademien der Wissenschaften, außeruniversitäre Forschungseinrichtungen der Max-Planck-Gesellschaft und der Leibniz-Gemeinschaft sowie einschlägige Fachverbände und -verbünde. Text+ umfasst außerdem führende Rechenzentren, die robuste und persistente Infrastrukturdienste für eine distribuierte Forschungsdateninfrastruktur bis hin zur Langzeitarchivierung absichern.

Die antragstellenden und beteiligten Institutionen unterhalten enge Verbindungen zu internationalen Forschungsdateninitiativen. Die Union der deutschen Akademien der Wissenschaften ist Mitglied von ALLEA, der Europäischen Föderation der Akademien der Wissenschaften. Die antragstellenden und beteiligten Institutionen sind aktive Teilnehmer in der weltweit agierenden Research Data Alliance (RDA). Teil von Text+ sind auch die nationalen Knotenpunkte in den europäischen Forschungsinfrastruktur-Konsortien CLARIN-ERIC und DARIAH-ERIC, die auch enge Beziehungen zu Forschungsinfrastrukturen außerhalb Europas aufgebaut haben. Die beiden ERICs und die in Text+ vertretenen Institutionen in Deutschland tragen durch Horizon2020-Projekte, z. B. EOSC-Pilot, EOSC-Hub und OpenAIRE-Advance, aktiv zur Gestaltung und Umsetzung der European Open Science Cloud (EOSC) bei, insbesondere über das Engagement in der Social Sciences and Humanities Cloud (SSHOC) als disziplinspezifische Anwendung von EOSC.

Die antragstellenden und beteiligten Institutionen sind in einschlägigen Ausschüssen nationaler und internationaler Organisationen aktiv in die Entwicklung, Verbreitung und Umsetzung von Standards für Forschungsdaten eingebunden. Dazu gehören insbesondere CIDOC, COAR, DataCite, Dublin Core Metadata Initiative, DIN, IIIF, ISO, das Kompetenzzentrum Interoperable Metadaten der DINI und das TEI-Konsortium. Damit leistet Text+ einen wichtigen Beitrag zur Standardisierung und Vernetzung von nationalen und internationalen Partnern und Infrastrukturen.

Schnittstellen zu der gesamten NFDI: Text+ hat die Berliner Erklärung (https://doi.org/10.5281/zenodo.3457213) zu NFDI-Querschnittsthemen und die Leipzig-Berlin-Erklärung (http://doi.org/10.5281/zenodo.3895208) zu NFDI-Querschnittsthemen der Infrastrukturentwicklung unterzeichnet. In enger Abstimmung mit dem NFDI-Direktorat wird Text+ in eine NFDI-weite Konsultation und Koordination zu den Research Data Commons eintreten. Die geisteswissenschaftlichen NFDI-Initiativen haben ihre Zusammenarbeit in einem Memorandum of Understanding (https://doi.org/10.5281/zenodo.3265763) spezifiziert. Darüber hinaus plant Text+ mit einzelnen Konsortien wie KonsortSWD und NFDI4Ing ausgewählte Themen gemeinsam zu bearbeiten.

Relevante Querschnittsthemen für Text+ und für die NFDI insgesamt und Beitrag von Text+: Text+ beinhaltet eine Vielzahl von auch für andere NFDI-Konsortien relevanten und nachnutzbaren Diensten, wie z. B. die Bereitstellung von Persistenten Identifikatoren (PIDs), eine AAI, Komponenten einer Metadaten-Infrastruktur und von Terminologie- sowie Text-Data-Mining-Diensten. Die Initiative wird sich an der gemeinsamen Entwicklung der Querschnittsthemen gemäß der Berliner und der Leipzig-Berlin-Erklärung beteiligen, u.a.: Ausbildung und Nutzenden-Support, Ethik- und Rechtsfragen, Infrastruktur für Daten-, Service- und Softwarequalität, Interoperabilitäts-Lösungen, Langzeitarchivierung.

Erwartungen an die NFDI-Konferenz: Text+ sieht die NFDI-Konferenz als Forum für den Informationsaustausch und die Vernetzung mit anderen NFDI Konsortien sowie für die Sichtung und Koordination von für Text+ relevanten Querschnittsthemen in Abstimmung mit dem NFDI-Direktorat.

Vorgesehene Mitglieder des Konsortiums (Co-Sprecherinnen/Co-Sprecher und die weiteren, beteiligten Institutionen):

Co-Sprecher/in	**Zugehörige Institution**
PD Dr. Alexander Geyken Co-Sprecher geyken@bbaw.de	Berlin-Brandenburgische Akademie der Wissenschaften Jägerstraße 22-23, 10117 Berlin
Prof. Dr. Wolfram Horstmann Co-Sprecher horstmann@sub.uni-goettingen.de	Georg-August-Universität Göttingen - Niedersächsische Staats- und Universitätsbibliothek Göttingen Platz der Göttinger Sieben 1, 37073 Göttingen
Dr. Peter Leinen Co-Sprecher p.leinen@dnb.de	Deutsche Nationalbibliothek Leipzig und Frankfurt am Main Adickesallee 1, 60322 Frankfurt am Main
Prof. Dr. Andreas Speer Co-Sprecher andreas.speer@uni-koeln.de	Nordrhein-Westfälische Akademie der Wissenschaften und Künste Palmenstraße 16, 40217 Düsseldorf

Vorgesehene weitere beteiligte Institutionen

Institution
Akademie der Wissenschaften zu Göttingen Theaterstraße 7, 37073 Göttingen
Akademie der Wissenschaften in Hamburg Edmund-Siemers-Allee 1, 20146 Hamburg
Akademie der Wissenschaften und der Literatur Mainz Geschwister-Scholl-Straße 2, 55131 Mainz
Albert-Ludwigs-Universität Freiburg: Englisches Seminar Rempartstr. 15, KG IV, Büro 4108, 79085 Freiburg
Eberhard-Karls-Universität Tübingen Geschwister-Scholl-Platz, 72074 Tübingen
Forschungszentrum Jülich GmbH Wilhelm-Johnen-Straße, 52425 Jülich
Georg-Eckert-Institut: Leibniz-Institut für internationale Schulbuchforschung Celler Str. 3, 38114 Braunschweig
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen Am Faßberg 11, 37077 Göttingen
Heidelberger Akademie der Wissenschaften Karlstraße 4, 69117 Heidelberg
Herzog August Bibliothek Wolfenbüttel Lessingpl. 1, 38304 Wolfenbüttel
Julius-Maximilians-Universität Würzburg: Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte Am Hubland, 97074 Würzburg
Karlsruher Institut für Technologie: Steinbuch Centre for Computing Kaiserstraße 12, 76131 Karlsruhe

Kooperation der Darmstädter wissenschaftlichen Einrichtungen: Hochschule Darmstadt: Arbeitsbereich Informationswissenschaft und Digitale Bibliothek Max-Planck-Straße 2, 64807 Dieburg; Technische Universität Darmstadt: Fachgebiet Computerphilologie und Mediävistik Dolivostraße 15, 64293 Darmstadt; Universitäts- und Landesbibliothek Darmstadt Magdalenenstr. 8, 64289 Darmstadt
Leopoldina: Nationale Akademie der Wissenschaften Jägerberg 1, 06108 Halle (Saale)
Ludwig-Maximilians-Universität München: Institut für Phonetik und Sprachverarbeitung Schellingstraße 3, 80799 München
Max Weber Stiftung Rheinallee 6, 53173 Bonn
Otto-Friedrich-Universität Bamberg: Lehrstuhl für Medieninformatik Fakultät WIAI, Lehrstuhl für Medieninformatik 96045 Bamberg
Sächsische Akademie der Wissenschaften zu Leipzig PF 100440, 04004 Leipzig
Salomon Ludwig Steinheim Institut für deutsch-jüdische Geschichte Edmund-Körner-Platz 2, 45127 Essen
Technische Universität Dresden, Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) 01062 Dresden
Universität Duisburg-Essen: Institut für Politikwissenschaft Forsthausweg 2, 47057 Duisburg
Universität Hamburg: Fakultät für Geisteswissenschaften Überseering 35, Postfach #15, 22297 Hamburg
Universität zu Köln: Cologne Centre for eHumanities (CCeH) Albertus-Magnus-Platz, 50923 Köln
Universität Paderborn: Detmold Hochschule für Musik Hornsche Straße 39, 32756 Detmold
Universität des Saarlandes Campus Geb. A2 2, 66123 Saarbrücken
Universität Trier: Trier Centre for Digital Humanities Universitätsring 15, 54286 Trier